

## REMARKS

Claims 12, 13, and 21-24 are pending. Claims 33-35 have been added. Support for the newly added claims is found in the specification and the claims as originally filed. The addition of new claims and the amendment of the existing claims do not affect inventorship.

### Specification: Incorporation by Reference

2. The Examiner states that the attempt to incorporate subject matter related to utilizing forcefield calculation into this application by referring to plurality of references cited in the paragraph bridging pages 14 and 15 is improper.

Applicant submits that the particular forcefield calculation used is not “essential” in that any forcefield calculation may be used. There is a commonality in all forcefield calculations (including all of those cited) in that all such calculations apply a forcefield of some sort to calculate the energies of chemical structures or groups of chemical structures. Applicant's invention is not limited to a particular forcefield calculation, but provides enabling disclosure by providing a plurality of references as examples of suitable forcefield calculations.

### **Claim Rejections – 35 USC § 112, second paragraph**

3. Claims 12, 21-24 are rejected under 35 USC § 112, second paragraph, as being indefinite for failing to particularly point out and distinctly claim the subject matter which applicant regards as the invention.

A. Claim 12, step (a): The Examiner states that the term “coordinates” is unclear. Applicants have amended this claim to more clearly identify the term “coordinates”. In this context, the term “coordinates” is well known in the art to reflect the structure coordinates (and thereby three dimensional) of a particular (target) protein. Support for the term “three dimensional” is found at page 8, line 23.

B. Claim 12, step (b): The Examiner states that the term “calculation” in regard to forcefield calculations is vague and indefinite. The method is not limited to a specific forcefield calculation, but is limited in that a forcefield calculation (as described on pages 14 and 15 of the specification) is used to generate a library of primary variants. The claim is limited to a forcefield as opposed to another type of function that could be applied to the target protein. Forcefield calculations are well known in the art of computational protein design (as exemplified by the list of references described in the specification), where protein sequence variants are scored by calculating their energies as chemical structures or as groups of chemical structures. And in response to Examiner's citation of Ulrich et al., the forcefield calculations of the present invention may be statistical or an average over physical terms.

- C. Claim 12, step (c): The Examiner states that there is no nexus between step (c) and other method steps. Applicant has eliminated this step and amended the last step (new step c) to clarify the nexus among the steps and to more closely conform to the claim as originally presented.
- D. Claims 21-24: The Examiner states that there is no nexus between the *in silico* method of claim 12 and the *in vitro* steps of claims 21-24. Applicant has amended claim 21 to clarify that the synthesis steps are done after the *in silico* method of claim 12.

#### **Claim Rejections – 35 USC § 101/112-1**

4. Claims 12, 21-24 are rejected under 35 USC § 101 because the claimed invention is not supported by either a specific asserted utility or a well established utility. The Office Action asserts that no substantial or credible asserted utility or well established utility has been disclosed. MPEP § 2701.02 IV states that “where a stated utility is not specific or substantial, a *prima facie* showing must establish that it is more likely than not that a person of ordinary skill in the art would not consider that any utility asserted by the applicant would be specific and substantial. The *prima facie* showing must contain the following elements: (A) An explanation that clearly sets forth the reasoning used in concluding that the asserted utility for the claimed invention is neither both specific and substantial nor well-established; (B) Support for factual findings relied upon in reaching this conclusion; and (C) An evaluation of all relevant evidence of record, including utilities taught in the closest prior art”.

Applicants respectfully request reconsideration of this rejection. Applicant’s believe that support is found in the specification itself as well as publications that used the methodology in the “real world”. As is stated in MPEP § 2107.02 III.A., the “Langer” test is to be used in a utility determination. As stated in Langer:

“As a matter of Patent Office practice, a specification which contains a disclosure of utility which corresponds in scope to the subject matter sought to be patented must be taken as sufficient to satisfy the utility requirement of § 101 for the entire claimed subject matter unless there is a reason for one skilled in the art to question the objective truth of the statement of utility or its scope.”

As further stated in MPEP § 2107.02 III.A.,

“[T]hus, Langer and subsequent cases **direct the Patent Office to presume that a statement of utility made by an applicant is true.** For obvious reasons of efficiency and in deference to an applicant's understanding of his or her invention, when a statement of utility is evaluated, Patent Office personnel should not begin an inquiry by questioning the truth of the statement of utility. Instead, **any inquiry must start by asking if there is any reason to question the truth of the statement of utility.** This

can be done by evaluating the logic of the statements made, taking into consideration any evidence cited by the applicant. **If the asserted utility is credible (i.e., believable based on the record or the nature of the invention), a rejection based on "lack of utility" is not appropriate.** Thus, Patent Office personnel should not begin an evaluation of utility by assuming that an asserted utility is likely to be false, based on the technical field of the invention or for other general reasons.

Compliance with § 101 is a question of fact. Thus, to overcome the presumption of truth that an assertion of utility by the applicant enjoys, Patent Office personnel must establish that it is more likely than not that one of ordinary skill in the art would doubt (i.e., "question") the truth of the statement of utility. To do this, Patent Office personnel must provide evidence sufficient to show that a person of ordinary skill in the art would consider the statement of asserted utility "false". A person of ordinary skill must have the benefit of both facts and reasoning in order to assess the truth of a statement. This means that if the applicant has presented facts that support the reasoning used in asserting a utility, Patent Office personnel must present countervailing facts and reasoning sufficient to establish that a person of ordinary skill would not believe the applicant's assertion of utility. The initial evidentiary standard used during evaluation of this question is a preponderance of the evidence (i.e., the totality of facts and reasoning suggest that it is more likely than not that the statement of the applicant is false)." (emphasis added).

As further outlined in MPEP § 2107.02. B:

"Where an applicant has specifically asserted that an invention has a particular utility, that assertion cannot simply be dismissed by Office personnel as being "wrong," even when there may be reason to believe that the assertion is not entirely accurate. Rather, Office personnel must determine if the assertion of utility is credible (i.e., **whether the assertion of utility is believable to a person of ordinary skill in the art based on the totality of evidence and reasoning provided**). **An assertion is credible unless (a) the logic underlying the assertion is seriously flawed, or (b) the facts upon which the assertion is based are inconsistent with the logic underlying the assertion.** Credibility as used in this context refers to the reliability of the statement based on the logic and facts that are offered by the applicant to support the assertion of utility."

It is respectfully submitted that the Examiner has not met this burden. A mere statement that Applicant has not met the burden does not shift the burden of utility to Applicant. The Examiner has not provided countervailing facts and reasoning. Moreover, Applicant has provided an explanation and evidence of utility by use of third party publications that assert the utility of the method of Applicant specifically (DeGrado) and the state of the art of computational design techniques generally (Saven).

Thus, Applicant has provided facts (2 publications by eminent persons in the field of computational design and evidence of “real world” commercial success) that are sufficient to establish that a person skilled in the art would believe the assertion of utility.

MPEP § 2017.02 IV states, “the PTO must do more than merely question operability – it must set forth factual reasons which would lead on skilled in the art to question the objective truth of the statement of operability”. Applicant respectfully submits that the Examiner has not provided factual reasons and further that Applicant has affirmatively provided evidence of the “objective truth of the statement of operability”.

Applicant's direct the Examiner's attention to In the article “Proteins from Scratch” (DeGrado, *Science* (1997), 278:80-81(Exhibit A). Dr. DeGrado states (in describing an earlier version of Applicant's methodology):

“Dahiyat...describe[s] a new approach that makes de novo protein design as easy as running a computer...Thus, the problem of de novo protein design reduced to two steps: selecting a desired tertiary structure and finding a sequence that would stabilize this fold. Dahiyat and Mayo have now mastered the second step with spectacular success. They have distilled the rules, insights and paradigms gleaned from two decades of experiments into a single computational algorithm...Thus the rules of ...computational methods for de novo design may now be sufficiently defined **to allow the engineering of a variety of proteins.**” (emphasis added)

The Saven publication (Exhibit B) shows that it is known in the art that combinatorial library generation has “real world use”:

“Not only can combinatorial methods be used for discovery but also, more deeply, they can inform our understanding of protein properties by generating and assaying whole ensembles of sequences. Traditionally, advances in structural biology have come from examining the structures of naturally occurring proteins, but with combinatorial experiments, **an enormous diversity of sequences can be generated at the control of the researcher**”.

In addition to these publications that demonstrate the state of the art and of one of ordinary skill in the art, The Examples in the specification establish the utility. Applicant displays “real world” use of computation design by citing **issued** patents that show the utility of this methodology. MPEP §2107.01 defines a “substantial utility” as a “real world” use. Applicant's have successfully used the claimed methods in the “real world” as shown in the publications cited as Exhibit B.

In addition, Applicant's have attached press releases showing that other companies have collaborated and are collaborating with Applicant using its computational methodologies, known commercially as Protein Design Automation® (PDA®).

As for “credibility”, the Examiner has not shown (a) the logic underlying the assertion is seriously flawed, or (b) the facts upon which the assertion is based are inconsistent with the logic underlying the assertion, as is required by the MPEP (see above), while Applicant has provided credible explanation and documentation.

Applicant recites rational methods to design proteins that utilize well known structural information and physico-chemical properties to provide novel proteins having designed characteristics. These methods are in contrast to prior art techniques such as alanine scanning and gene “shuffling” which are not rational and require extensive experimentation to determine proteins’ properties and sequence. The methodology has incorporated this information to address the “complex nature” of secondary structure. The method yields “real world” results, not “prophetic or expedient statements”. This is shown in the examples in the specification, the “real world” examples as described in scientific publications and in the opinion of those skilled in the art.

The Examiner states in the Office Action at page 5 that, “creating a library for further screening or testing is not a utility for a method”. In addition to the utility and “real world” uses described above, Applicant directs the Examiner to the specification at page 1, lines 21-23, which states, “computational methods can be used to screen enormous sequence libraries (up to  $10^{80}$  in a single calculation) overcoming the key limitation of experimental library screening methods”; page 5, lines 14-19, which states that the “invention can be used to prescreen libraries based on known scaffold proteins...using computational methods, the percentage of useful variants in a given variant set size can increase, and the required experimental outlay is decreased.”; page 6, lines 28-33, which states,

“virtual libraries of protein sequences can be generated that are vastly larger than experimental libraries...sequences can be screened and those that meet design criteria...can be readily selected. An experimental library consisting of the favorable candidates found in the virtual library screening can then be generated, resulting in a much more efficient use of the experimental library and overcoming the limitations of random protein libraries.”

Page 7, lines 1-16 describes the advantages of the computational methodologies as providing a

“list of sequence candidates that are favored to meet design criteria; it also shows which positions in the sequences are readily changed and which positions are unlikely to change without disrupting protein stability and function. The diversity of amino acids at these positions can be limited to those that are compatible with these positions. Thus, the number of wasted sequences produced in the experimental library is reduced, thereby increasing the probability of success in finding sequences with useful properties. In addition... greater diversity of protein sequences can be screened (i.e. a larger sampling of sequence space), leading to greater improvements in protein function.

Further, fewer mutants need to be tested experimentally to screen a given library size, reducing the cost and difficulty of protein engineering.”

Thus, prima facie shifted to the Examiner. The Examiner analogizes a library to a composition of matter, which has to undergo screening to isolate and identify a product, citing Brenner v. Manson, 148 USPQ 689 (1966) (“Brenner”).

Applicants are specifically claiming a method of generating a secondary library, not a “library” per se, nor a composition of matter in the instant application. Thus, the analogy to Brenner that the Examiner makes is not analogous to the claims in the instant application. In addition, Applicants respectfully disagree with the analogy to Brenner because the protein variants to be screened by the method of the present invention, find utility in their respective fields. For example, for purposes of the present invention, it does not matter what the class of proteins are. The method of the claimed invention, screens for useful variants having desired protein characteristics. See for example, Specification at page 4, lines 25-30 and page 34, lines 22 to page 35, line 12. For example, the variants produced from the method of the present invention may find use as therapeutic proteins. See Specification beginning at page 34, lines 22, ending on page 35, line 12.

Because the method starts with a target protein with known three dimensional coordinates, the structure of the variant protein sequences of the secondary library are known. Thus, the Examiner’s statement that the “secondary library of as yet undefined structure” is incorrect.

The “actual and specific significance...attributed to the secondary library” (see Office Action, page 6, first full paragraph) is that the skilled artisan is starting with a known target protein and inputting three dimensional coordinates (that is, the structure and sequence of the starting protein is known), then design criteria are applied to enhance the physico-chemical properties of the target protein. For example, a skilled artisan could decide to enhance the stability of a target protein. By applying the methods described in the specification, the resulting secondary library would have sequences with enhanced stability as compared to the target protein. Thus the artisan would not be required to perform additional experimentation to determine how to use the generated secondary library. The usefulness of the library was established by the design criteria. The real world utility is that the library provides a plurality of sequences that meet the design criteria. The advantage of Applicant’s method is that the library generated can be extremely diverse – exploring space not readily explorable by random techniques – or narrowly focused to identify a very specific subset of sequences that meet the design criteria.

The “potential pharmacological utility” is known and disclosed since the design criteria establishes the utility. The correlation and relationship between the pharmacological sequence and the disease is already established, the goal of the method is to enhance the property of the sequence to treat the disease.

Moreover, the existence of the methodology per se provides a utility since it provides libraries that are different from the target protein, said library members possessing designed properties. Both Saven and DeGrado have found this to be a pioneering advance in the art of protein design.

5. The arguments made above with respect to 35 USC §101 are equally applicable to the rejection under 35 USC §112, first paragraph. The techniques described in the recited methods have a specific and well-established utility, and one skilled in the art would know how to use the claimed invention, particularly as demonstrated in the patents and scientific articles, as well as the commercial success, discussed above.

Thus, the burden is shifted to the Examiner. The Examiner analogizes a library to a composition of matter, which has to undergo screening to isolate and identify a product, citing *Brenner v. Manson*, 148 USPQ 689 (1966). Applicants respectfully disagree because the invention as claimed is in fact a method for generating libraries. Although the examiner describes the secondary libraries as presently undefined, the method for generating them is fully enabled by the specification (see the discussion below regarding written description). The basis for this is that Applicant's are claiming a method of generating a secondary library. The library generated will necessarily vary with the particular target protein identified, as well as the use of the different parameters of the method.

Thus, the discussions above regarding examples of actual utility by Applicant, as well as recognition to those skilled in the art of protein design and combinatorial library generation, meets the utility requirement under 35 USC § 101. It is submitted that the present invention has utility under §101 Applicants respectfully request that the rejection be withdrawn.

#### **Claim Rejections – 35 USC § 112, first paragraph (Written Description)**

6. Claims 12, and 21-24 are rejected under 35 USC § 112, first paragraph, as failing to comply with the written description requirement. The Examiner states that there is no description in the specification of the method that includes all of the method steps as claimed.

It is noted that the claims have been modified to address the previous and current Examiner's concerns about the individual steps. Moreover, there is more than adequate support for each of the steps in the specification. The specification describes the flexibility of Applicant's

design approach in that many different functions and steps may be combined to generate secondary libraries with enhanced diversity. Thus, the specification does describe alternative approaches to protein modeling, but it additionally provides various combinations of these various approaches to generate secondary or tertiary libraries. The purpose of the various combinations is to allow the skilled artisan flexibility in establishing design criteria to generate a library that meets the artisan's design goals.

The Examiner is directed to the Example 1, starting at page 61, line 7 of the specification. Starting at line 8, a  $\beta$  lactamase molecule is pre-screened using PDA® technology. It is noted that the structure of this molecule is known, these structure coordinates are input to the computer and certain of the residues are assigned for sequence variation analysis, i.e. floated. Thus, step a) of the recited independent claims is specifically described here (e.g., "inputting the three dimensional coordinates of said target protein into a computer"). A forcefield calculation is applied (starting at page 61, line 28 of the specification) to generate a primary library (see Table 3 on page 62). Thus, step b) of the recited independent claims is specifically described here (e.g., "utilizing a forcefield calculation to generate a primary library comprising a plurality of primary variant amino acid residues at primary variant positions").

The amino acid variants contained in this primary library are recombined in Table 4 (on page 63) to generate a secondary library as in step c) of the independent claims (e.g., "combining a plurality of said primary variant amino acid residues from step b) to generate a secondary library of secondary variant proteins").

The library is then optionally synthesized (starting at page 63 through page 65, line 25) resulting in a variant that improved enzyme function. In this library, the mutant sequences are "made in the desired proportions as shown in Table 4... and equal molar concentrations of the oligonucleotides were pooled (see claims 21-23). Also, in the paragraph starting on page 64, line 11, the oligonucleotides are added according to their frequency (claim 24).

Example 2 provides the secondary library of a xylanase. Again this molecule whose three dimensional coordinates are known (see page 65, line 31 through page 66, line 3 and Figure 2), therefore, step a) of the recited independent claims is specifically described here (e.g., "inputting the three dimensional coordinates of said target protein into a computer"). A forcefield was applied to generate a primary library starting at page 66, line 4 and Table 1 (also page 66). Thus, step b) of the recited independent claims is specifically described here (e.g., "utilizing a forcefield calculation to generate a primary library comprising a plurality of primary variant amino acid residues at primary variant positions").



Table 2 (on page 67) shows the ability of the method to recombine variants for a secondary library that is smaller and more focused, as in step c) of the independent claims (e.g., “combining a plurality of said primary variant amino acid residues from step b) to generate a secondary library of secondary variant proteins”).

Additional information on the computational design methodologies used to generate libraries of novel xylanase variants may be found in US 6,627,186, owned by the same assignee as this application.

Thus, the disclosure and combination of the various computational techniques in the specification does provide an adequate written description and enables a method for computationally generating secondary libraries comprising variant sequences in which the starting target protein structure (*i.e.* scaffold) can be any protein for which a three dimensional structure is known or can be generated.

As discussed above, the specification does describe every element of the claimed invention. The specification is set up to emphasize the flexibility of the methodology, in that many different steps may be combined to generate secondary or tertiary libraries. The size and diversity is dependent upon the goals of the skilled artisan. The specification provides a written description that is in sufficient detail that a skilled artisan could combine the various elements to generate the described method. General support for this method is found at page 8, lines 8-12 which describes the general nature of the method, that is, the ability to combine a primary library to generate a secondary library.

In the instant case, in step a) of claim 12, three dimensional coordinates of a target protein are input into a computer. Support for step a) is found at page 8, line 23.

Then a forcefield is applied to a primary library and then combined to generate a secondary library. Support for generating a primary library via a forcefield is found at page 14, line 23 through page 15, line 14 and at page 22, lines 1 through 22. At page 23, lines 16-20, the specification states that,

“[A]s will be appreciated by those in the art, once... [a] set of sequences is generated (e.g., step b) of claim 12)...a variety of sequence space sampling methods can be done...That is, once a...set of sequences is generated, preferred methods utilize sampling techniques to allow the generation of additional, related sequences for testing. These sampling methods can include...recombinations of one or more sequences (e.g., new step c) of claim 12).

At page 24, lines 28-29, the specification states,

“these sampling methods can be used to further process a secondary library to generate additional secondary libraries.”

Page 27, lines 20-26 of the specification states,

“These primary library positions can then be recombined to form a secondary library (e.g., step c) of Claim 12)...The formation of the secondary library using this method may be done in two general ways; either all variable positions are allowed to be any amino acid, or subsets of amino acids are allowed for each position.”

Thus, the above section of the specification provides support for the term “plurality” in step c).

Further support is found at page 30, lines 19-27 of the specification, where it is stated that,

“secondary libraries can be generated... computationally, wherein the primary library is further computationally manipulated, for example...by recombining portions of the sequences containing one or more variant position... This computationally-derived secondary library can be experimentally generated by synthesizing the library members...”

This section of the specification provides support for step c) of claim 12 and for the synthesis claims 21-24. Specific support for claims 21-24 is found at page 31, line 8 through page 34, line 8 and at page 40, lines 25 through 35.

The articles, press releases, patents and patent applications discussed above with respect to the 101 rejection support the enablement of the methods disclosed in the pending claims.

Accordingly, Applicants respectfully submit that the specification fully enables the present claims, and respectfully request withdrawal of the rejection under 35 U.S.C. § 112, first paragraph.

### **Double Patenting**

7. Claims 12 and 21-24 are rejected under the judicially created doctrine of obviousness-type double patenting as being unpatentable over claims 1-8 of US Patent 6,403,312.

Applicants have attached a terminal disclaimer to overcome this rejection. Applicant also confirms that the instant application and US 6,403,312 are assigned to the same assignee.

8. Claims 12 and 21-24 are provisionally rejected under the judicially created doctrine of obviousness-type double patenting as being unpatentable over claims 19-29 of co-pending application no. 09/927790.

Applicant respectfully requests that the claim scope be reevaluated once the claims of both applications are in condition for allowance.

### **Claim Rejections – 35 USC § 102 and 103**

Claim 12 is rejected under 35 USC § 102(e) as being anticipated by Lacroix et al (US 2002/0072864, filing date 08/31/1999) (Lacroix). It is noted that Lacroix has an effective filing date less than 7 months before the effective date of the instant application.

"A claim is anticipated only if each and every element as set forth in the claim is found, either expressly or inherently described, in a single prior art reference." *Verdegaal Bros. v. Union Oil Co. of California*, 814 F.2d 628, 631, 2 USPQ2d 1051, 1053 (Fed. Cir. 1987). "The identical invention must be shown in as complete detail as is contained in the ... claim." *Richardson v. Suzuki Motor Co.*, 868 F.2d 1226, 1236, 9 USPQ2d 1913, 1920 (Fed. Cir. 1989)."

Lacroix teaches a method for choosing a set of amino acids in a target protein by identifying at least one substitute for each amino acid position in the set; determining at least one conformer for each substitute, substituting coordinates of each conformer for the coordinates of the positions in the target protein; minimizing the value of a calculated solution score by adjusting the geometry of the conformer to obtain a "solution structure"; and determining whether the "solution structure" has a score that is lower than a threshold value.

More specifically, Lacroix discloses energy calculations, followed by dead end elimination to reduce rotamer complexity, and finally applies mean-field optimization iteratively to generate rotamer probabilities (as in section 5.9.2 paragraph [0164]). Lacroix further discusses output of optimal sequence results based on the former. However, the connection between the two is not clearly described, nor suggested. The rotamer probabilities disclosed in section 5.9.2 paragraph [0164]) specifically refer to the mean-field weights generated within the mean field method, describing what happens in the dynamic cycling between rotamer energies and rotamer probabilities, as is typical for mean-field optimization methods. Nowhere is it taught or disclosed how to use these probabilities directly to generate more than one sequence.

Assuming *arguendo*, that Lacroix anticipates the connection, the end result would be that Lacroix is just making a primary library directly from the rotamer probability is not a secondary library.

The specification at page 10, line 17 through page 26, line 26 defines a primary library and the various techniques for generating a primary library. A secondary library is defined in the specification at page 26, line 27 through page 31, line 7. To summarize these sections, a primary library is a set of optimized sequences and a secondary library is preferably generated from a primary library. The secondary library can be a subset, or contain new members not found in the primary library. Variant positions and/or amino acid residues in the variant positions can be combined or sampled in any number of ways to form a new library that exploits the sequence variations found in the primary library.

Lacroix actually teaches away from using the probabilities directly in Section 5.11. at paragraph 0173 as follows:

"After the dead-end elimination procedure of the preferred embodiment of the present invention, **many sequences remain; nevertheless, the subsequent steps (see Mean Field Theory and Refinement, above) ensure that only those conformations and sequences that satisfy predetermined energy thresholds finally surface as candidates for the target structure.** The preferred embodiment of the present invention can produce either detailed or limited outputs, depending on the size of the sampled sequence space. In one embodiment, the **output is a simple list of sequences and scores** (evaluated using the scoring function) that can be sorted according to the calculated potential energy so that the **lowest energy sequences may be readily identified.** In another embodiment, a more complete output presents a numerical profile of the energy for each sequence produced. The program is also capable of producing a coordinate file (in PDB format) with the structure of the protein having a mutated sequence. **If mean field sampling is performed, both the PDB-file and detailed energy outputs correspond to the combination of most probable rotamers**" (emphasis added).

In summary, Lacroix does not disclose all of the steps of Applicant's methodology. There is no suggestion, teaching or disclosure of using a forcefield calculation **and** a recombination of such modeled proteins to generate a secondary library of secondary sequences. Applicants note that when Lacroix refers to a "library", Lacroix means a set of rotamers to be substituted at a particular amino acid position (rotamer library is a term of art for a generic collection of rotamer states).

LaCroix is generally looking for a solution structure, or a small collection of low energy solution structures. Lacroix's only example lists the equivalent of a **primary** library at paragraph 0196:

"Sample sequences along with their solution scores as output from the program are as follows: 5 CORE SURFACE VAVMLLVVV -76.6 (Wild Type) VNDR -6.6 (Wild Type) LVIVLLVIV -81.8 VGSK -28.0 VVILLVIV -81.8 IVIILLVVV -81.9 LIIVLLVIV -82.0 IVVILLVIV -82.8 IIVLLVIV -82.8 IVLILLVIV -82.9 LVIILLVIV -83.2 IVIILLVIV -84.4 "

Thus, there is no generation of a secondary library in Lacroix.

Claims 21-24 are rejected under 35 USC § 103(a) as obvious over Lacroix for teaching synthesizing candidate structures. Applicant confirms that the subject matter of the various claims was and is commonly owned.

To establish a *prima facie* case of obviousness, three basic criteria must be met: 1) suggestion or motivation, either in the references themselves or in the knowledge generally available to one of ordinary skill in the art to modify or combine reference teachings; 2) there

must be a reasonable expectation of success; and 3) the prior art reference must teach or suggest all the claim limitations. (See MPEP §2142).

As discussed above, Lacroix does not anticipate claim 12 from which claims 21-24 depend since Lacroix never generates a secondary library, nor does Lacroix teach or suggest making a secondary library. Therefore, the Examiner has not established a prima facie case of obviousness under § 103.

Lacroix states at paragraph 0179 that,

"This system, when operated in a laboratory environment can provide an efficient and useful method of directing experimental efforts towards engineering sequence variations in a target macromolecule. Said system, being capable of quantifying the potency of a plurality of sequences and **thereby selecting a small number which would be worthy of synthesis**, can operate in tandem with experiment to optimize properties of interest of the target macromolecule." (emphasis added)

Lacroix was merely synthesizing a small number of variant sequences that would be "worthy of synthesis". Lacroix is not even synthesizing a library of variants, Lacroix is just making either a "solution structure" or a couple of "solution structures". Thus, Lacroix does not even disclose synthesizing a primary library, let alone a secondary library as Applicant has specifically recited in Claim 12 from which claims 21-24 depend.

Further, the Examiner acknowledges Lacroix does not teach the PCR methodology as specifically recited by Applicant in these claims. The PCR methodology of Applicant is used with the computational methodology recited in claim 12 to generate libraries of differing diversity. Lacroix does not contemplate or suggest the use of the PCR tools to create additional diversity but merely synthesizes the "solution structure", not a library with varying diversity, nor a modeled secondary library, nor a secondary library at all.

Thus, while Lacroix may teach synthesis *per se*, Lacroix does not teach the specifically recited methods of Applicant. Applicant is not claiming synthesis techniques *per se*, but synthesis of at least a portion of the secondary library of modeled sequences by these techniques. These techniques are defined only in the context of Applicant's defined method, not a synthesis *per se*.

Further, Applicant has recited claims that are directed to further diversifying the secondary library by use of the various synthesis techniques described in claims 21-24, not just merely making a variant or two as Lacroix does. One of the goals of the various PCR techniques is to enhance diversity or to generate a more focused library. As discussed earlier with respect to the written description rejection, Example 1 ( $\beta$  lactamase) shows the diversity and focus in using the PCR techniques defined in Claims 21-24.

In conclusion, the instant invention and claims are not anticipated by Lacroix because Lacroix does not generate secondary libraries. Further, as acknowledged by the Examiner, the Lacroix reference does not teach the techniques recited in claims 21-24. Finally, since there is no suggestion to generate a library, let alone a secondary library in Lacroix, and the fact that Lacroix merely synthesizes a "solution structure" or two, there is no *prima facie* case of obviousness established.

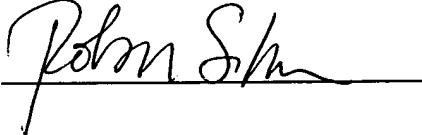
The Applicants submit that in light of the above arguments, the claims are now in condition for allowance and an early notification of such is respectfully solicited. The Examiner is invited to contact the undersigned at (415) 781-1989 if any issues may be resolved in that manner.

Dated: May 9, 2005

Four Embarcadero Center  
Suite 3400  
San Francisco, California 94111-4187  
Telephone: (415) 781-1989  
Fax No. (415) 398-3249

Respectfully submitted,

DORSEY & WHITNEY LLP

By: 

Robin M. Silva, Reg. No. 38,304  
Filed under 37 C.F.R. § 1.34

# Proteins from Scratch

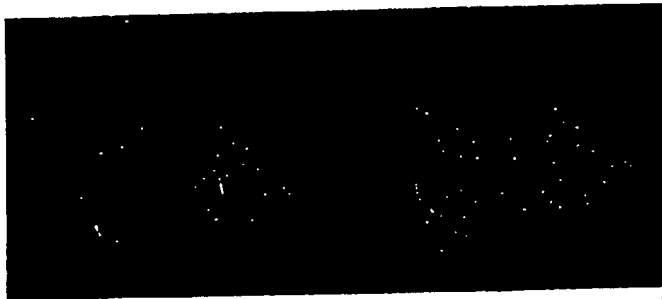
William F. DeGrado

Not long ago, it seemed inconceivable that proteins could be designed from scratch. Because each protein sequence has an astronomical number of potential conformations, it appeared that only an experimentalist with the evolutionary life span of Mother Nature could design a sequence capable of folding into a single, well-defined three-dimensional structure. But now, on page 82 of this issue, Dahiyat and Mayo (1) describe a new approach that makes de novo protein design as easy as running a computer program. Well almost...

The intellectual roots of this new work go back to the early 1980s when protein engineers first thought about designing proteins (2). At that point, the prediction of a protein's three-dimensional structure from its sequence alone seemed a difficult proposition. However, they opined that the inverse problem—designing an amino acid sequence capable of assuming a desired three-dimensional structure—would be a more tractable problem, because one could "over-engineer" the system to favor the desired folding pattern. Thus, the problem of de novo protein design reduced to two steps: selecting a desired tertiary structure and finding a sequence that would stabilize this fold. Dahiyat and Mayo have now mastered the second step with spectacular success. They have distilled the rules, insights, and paradigms gleaned from two decades of experiments (3) into a single computational algorithm that predicts an optimal sequence for a given fold. Further, when put to the test the algorithm actually predicted a sequence that folded into the desired three-dimensional structure. Thus, the rules of protein folding and computational methods for de novo design may now be sufficiently defined to allow the engineering of a variety of proteins.

Dahiyat and Mayo's program divides the interactions that stabilize protein structures

into three categories: interactions of side chains that are exposed to solvent, of side chains buried in the protein interior, and of parts of the protein that occupy more interfacial positions. Exposed residues contribute to stability, primarily through conformational preferences and weakly attractive, solvent-exposed polar interactions (4). The burial of hydrophobic residues in the well-packed in-



**Better than the real thing.** The natural zinc finger protein Zif268 (left) is stabilized in part by a core of hydrophobic (green) side chains and metal-chelating side chains (red). In the designed protein FSD-1 (right), the Zif268 core is retained but the metal-chelating His residues and one of the Cys residues of Zif268 are converted to hydrophobic Phe and Ala residues, thereby extending the hydrophobic core. The fourth metal ligand Cys<sup>6</sup> is converted to a Lys residue. The apolar portion of this interfacial residue shields the hydrophobic core, whereas its ammonium group is exposed to solvent. The helix is also stabilized by an N-capping interaction (19), which presumably also stabilizes the structure.

terior of a protein provides an even more powerful driving force for folding. The side chains in the interior of a protein adopt unique conformations, the prediction of which is a large combinatorial problem.

One important simplifying assumption arose from the early work of Jainin *et al.* (5), who showed that each individual side chain can adopt a limited number of low-energy conformations (named rotamers), reducing the number of probable conformers available to a protein. This work was subsequently extended to the design of proteins containing only the most favorable rotamers (6). Although the side chains in natural proteins deviate from ideality in a few cases (complicating the prediction of the structures of natural proteins), these deviations need not be considered in the design of idealized proteins. Thus, various algorithms have been developed to examine all possible hydrophobic residues in all possible rotameric states, to find combinations that efficiently fill the interior of a protein. A complementary ap-

proach uses genetic methods to exhaustively search for sequences capable of filling a protein core (7), and this work has been adapted for the de novo design of proteins (8).

Interfacial residues are also quite important for protein stability (9, 10). They are often amphiphilic (for example, Lys, Arg, and Tyr) and their apolar atoms can cap the hydrophobic core, while their polar groups engage in electrostatic and hydrogen-bonded interactions.

Until recently, protein designers have frequently concentrated on quantifying the energetics associated with just one of these three types of interactions (3). However, de novo design is best approached by simultaneously considering all of the side chains in the protein—unfortunately, a very high-order combinatorial problem. For instance, the volume available to the interior side chains depends on the nature and conformation of the residues at the interfacial positions and vice versa. Dahiyat and Mayo assumed that each of these three features had been adequately quantitated to provide a useful empirical energy function for protein design. Their program combines a number of features taken from earlier potential functions and includes a penalty for exposing hydrophobic groups to solvent. Another essential innovation included in their program is an implementation of the Dead-End Elimination theorem, to efficiently search through sequence and side chain rotamer space.

Dahiyat and Mayo's target fold is a zinc finger, a motif with a well-established history in protein structure prediction and design. In an early, prescient paper, Berg correctly inferred that this His<sup>2</sup>Cys<sup>2</sup> Zn-binding motif must feature a  $\beta$ - $\beta$ - $\alpha$  fold that would position the ligating groups in a tetrahedral array around the bound Zn(II) (11). Favorable metal ion-ligand interactions together with a small apolar core help stabilize the three-dimensional structure of this compact fold. More recently, Imperiali and co-workers have designed a peptide that folded into this motif, even in the absence of metal ions (12). The design included a D-amino acid to stabilize a type II' turn, and a large, rigid tricyclic side chain that may help consolidate the hydrophobic core. This work was particularly ex-

An enhanced version of this Perspective with links to additional resources is available for Science Online subscribers at [www.sciencemag.org](http://www.sciencemag.org)

The author is in the Department of Biochemistry and Biophysics, University of Pennsylvania School of Medicine, Philadelphia, PA 19104-6059, USA. E-mail: [wdegrado@mail.med.upenn.edu](mailto:wdegrado@mail.med.upenn.edu)

citing because, before their studies, it was not expected that sequences as short as 25 residues in length could fold into stable tertiary structures.

Now, Dahiyat and Mayo take these studies one step further through the design of a sequence composed of only natural amino acids that adopts the zinc finger motif. As input to their program, they introduced the coordinates of the backbone atoms from the crystal structure of the second domain of the zinc finger protein Zif268. The program then evaluated a total of  $10^{62}$  possible side chain-rotamer combinations to find a sequence capable of stabilizing this fold without a bound metal ion. The resulting protein sequence shares a small hydrophobic core with its predecessor from Zif268. However, in the newly designed protein FSD-1 the core is enlarged through the addition of hydrophobic residues that fill the space vacated by the removal of the metal-binding site (see the figure). This increase in the size of the hydrophobic core together with the enhancements in the propensity for forming the appropriate secondary structure provide an adequate driving force for folding. The designed miniprotein actually folds into the desired structure as assessed by nuclear magnetic resonance spectroscopy, and the observed structure closely resembles the three-dimensional structure of Zif268.

Because of its small size, the protein is marginally stable. A Van't Hoff analysis of the thermal unfolding curve gives a change in the enthalpy ( $\Delta H_{UH}$ ) of approximately  $-10$  kcal/mol, and indicates that the protein is about 90 to 95% folded at low temperatures (13). The small value  $\Delta H_{UH}$  and the lack of strong cooperativity in the unfolding transition are expected for a native-like protein of this very small size (14). Thus, FSD-1 is the smallest protein known to be capable of folding into a unique structure without the thermodynamic assistance of disulfides, metal ions, or other subunits. This important accomplishment illustrates the impressive ability of Dahiyat and Mayo's program to design highly optimized sequences.

This new achievement caps a banner year for de novo protein design. Earlier, Regan (15) answered the challenge of changing a protein's tertiary structure by altering no more than 50% of its sequence. And although Dahiyat and Mayo have demonstrated that the stabilizing metal-binding site is not necessary in their system, Caradonna, Hellinga, and co-workers (16) have made impressive progress in automating the introduction of functional metal-binding sites into the three-dimensional structures of natural proteins. Further, other workers (17) have used less automated approaches to successfully introduce functionally and spectroscopically interesting metal-binding sites into de novo designed proteins.

To date, the most computationally intensive protein design problems have been the redesign of natural proteins of known three-dimensional structure. But the new automated approaches open the door to the de novo design of structures with entirely novel backbone conformations. It will be interesting to see if Dahiyat and Mayo's approach of designing an optimal sequence for a given fold is sufficient, or if it will be necessary also to destabilize alternate possible folds. Indeed, when using an earlier version of their algorithm to repack the interior of the coiled coil from GCN4, they had to retain the identity of a buried Asn residue from the wild-type protein. Although the inclusion of this Asn actually destabilized the desired fold, it was nevertheless essential to avoid the formation of alternate, unwanted conformers (18). The ability to ask such focused questions will reveal much about how natural proteins adopt their folded conformations while simultaneously allowing the design of entirely new polymers for applications ranging from catalysis to pharmaceuticals.

#### References and Notes

1. B. I. Dahiyat and S. L. Mayo, *Science* **278**, 82.
2. K. E. Drexler, *Proc. Natl. Acad. Sci. U.S.A.* **78**, 5275 (1981); C. Pabo, *Nature* **301**, 200 (1983).
3. W. F. DeGrado, Z. R. Wasserman, J. D. Lear, *Science* **243**, 622 (1989); J. W. Bryson *et al.*, *ibid.* **270**, 935 (1995); M. H. J. Cordes, A. R. Davidson, R. T. Sauer, *Curr. Opin. Struct. Biol.* **6**, 3 (1996).
4. R. Munoz and L. Serrano, *Proteins* **20**, 301 (1994); C. A. Kim and J. M. Berg, *Nature* **362**, 267 (1993); D. L. Minor and P. S. Kim, *ibid.* **367**, 660 (1994); C. K. Smith, J. M. Withka, L. Regan, *Biochemistry* **33**, 5510 (1994).
5. J. Janin, S. Wodak, M. Levitt, B. Maigret, *J. Mol. Biol.* **125**, 37 (1978).
6. J. W. Ponder and F. M. Richards, *ibid.* **193**, 775 (1987); J. R. Desjarlais and T. M. Handel, *Protein Sci.* **4**, 2006 (1995); X. Jing, E. J. Bishop, R. S. Farid, *J. Am. Chem. Soc.* **119**, 838 (1997).
7. J. U. Bowie, J. F. Reidhaar-Olson, W. A. Lim, R. T. Sauer, *Science* **247**, 1306 (1990).
8. S. Kamlekar, J. M. Schiffer, H. Xiong, J. M. Babik, M. H. Hecht, *ibid.* **262**, 1680 (1993).
9. K. J. Lumb and P. S. Kim, *ibid.* **271**, 1137 (1996); Y. Yu, O. D. Monera, R. S. Hodges, P. L. Privalov, *J. Mol. Biol.* **255**, 367, (1996).
10. A. C. Braisted and J. A. Wells, *Proc. Natl. Acad. Sci. U.S.A.* **93**, 5688 (1996).
11. J. M. Berg, *ibid.* **85**, 99 (1988).
12. M. D. Struthers, R. P. Cheng, B. Imperiali, *Science* **271**, 342 (1996).
13. This Van't Hoff analysis of the protein is approximate because of the lack of definition of the pre- and posttransition baselines.
14. P. Alexander, S. Fahnestock, T. Lee, J. Orban, P. Bryn, *Biochemistry* **31**, 3597 (1992).
15. S. Dalal, S. Balasubramanian, L. Regan, *Nat. Struct. Biol.* **4**, 548 (1997).
16. A. Pinto, H. W. Hellinga, J. P. Caradonna, *Proc. Natl. Acad. Sci. U.S.A.* **94**, 5562 (1997); C. Coldren, H. W. Hellinga, J. P. Caradonna, *ibid.*, p. 6635.
17. B. R. Gibney, S. E. Mulholland, F. Rabanal, P. L. Dutton, *ibid.* **93**, 15041 (1996); M. P. Scott, J. Biggins, *Protein Sci.* **6**, 340 (1997); P. A. Arnold, W. R. Shelton, D. R. Benson, *J. Am. Chem. Soc.* **119**, 3181 (1997); G. R. Dieckman *et al.*, *ibid.*, p. 6195.
18. P. B. Harbury, T. Zhang, P. S. Kim, T. Alber, *Science* **262**, 1401 (1993); K. J. Lumb and P. S. Kim, *Biochemistry* **34**, 8642 (1995).
19. L. G. Presta and G. D. Rose, *Science* **240**, 1632 (1988); J. S. Richardson and D. C. Richardson, *ibid.*, p. 1648.



## Combinatorial protein design

Jeffery G Saven

Combinatorial protein libraries permit the examination of a wide range of sequences. Such methods are being used for *de novo* design and to investigate the determinants of protein folding. The exponentially large number of possible sequences, however, necessitates restrictions on the diversity of sequences in a combinatorial library. Recently, progress has been made in developing theoretical tools to bias and characterize the ensemble of sequences that fold into a given structure — tools that can be applied to the design and interpretation of combinatorial experiments.

### Addresses

Department of Chemistry, University of Pennsylvania, 231 South 34 Street, Philadelphia, Pennsylvania 19104, USA;  
e-mail: saven@sas.upenn.edu

Current Opinion in Structural Biology 2002, 12:453–458

0959-440X/02/\$ — see front matter

© 2002 Elsevier Science Ltd. All rights reserved.

### Introduction

The discovery and design of novel proteins can lead to new, potentially practical proteins and can also enhance our understanding of protein biochemistry. Designing well-structured, soluble proteins is difficult, however, because of their complexity. Such proteins are large (tens to hundreds of amino acid residues) and have many variables that specify the folded state, including sequence, backbone topology and sidechain conformation. Design involves identifying those sequences that fold into a given structure from a huge ensemble of possible sequences. This search is aided, in part, by the large degree of consistency seen in folded proteins. On average, a folded structure is well packed, hydrophobic residues are sequestered from solvent and most potential hydrogen bond interactions are satisfied. This consistency, however, is often complex, may have little simplifying symmetry and involves predominantly noncovalent interactions. Such interactions are some of the most difficult to accurately quantify. As such, estimating the free energies associated with mutation or structural ordering remains a subtle area of computational research. Nonetheless, many molecular potentials do contain a 'best parameterization' of many of the interatomic interactions and forces that we know are important for stabilizing proteins. In some cases, such potentials have been used with striking success in protein design [1\*\*]. Given that these potentials are necessarily approximate, however, one promising approach is to use the partial information contained in these functions in a probabilistic manner. A probabilistic or statistical approach is also appropriate for characterizing the full variability of sequences that fold to a common structure, because there are likely to be an enormous number of such sequences. Such statistical methods can be applied in 'shotgun' approaches to *de novo* protein design. Combinatorial experiments create and assay

many sequences in order to overcome shortcomings in our understanding of folding or other molecular properties. Even though combinatorial methods can address large numbers of sequences ( $10^4$ – $10^{12}$ ), these numbers are still infinitesimal in comparison to the numbers of possible sequences (e.g.  $20^{100} \approx 10^{130}$  for a 100-residue protein). Thus, methods for winnowing and focusing sequence space are a vital component of combinatorial protein design. Herein, I briefly discuss combinatorial methods for full sequence design. I also review recent theoretical developments in characterizing sequence ensembles — developments that can be applied to the design and interpretation of combinatorial experiments.

### Directed protein design

There has been much effort — and success — in developing computational methods for 'directed' protein design. By 'directed protein design', I mean the identification of a sequence (or a small set of sequences) that is likely to fold into a predetermined backbone structure. Each such sequence can then be synthesized to confirm its folded structure and other molecular properties. Early efforts in design identified proteins with substantial order, but not necessarily well-defined tertiary structure [2]. Because an enormous number of sequences are possible even for small proteins (<50 residues), computational methods have dramatically accelerated successful design. Typically, such methods are implemented as an optimization process, whereby amino acid identity and sidechain conformation are varied in order to optimize a scoring function that quantifies sequence/structure compatibility. Exhaustive searching of all  $m^N$  possible sequences (where  $m$  is the number of different amino acid types or 'states' per residue and  $N$  is the number of residues in a target protein structure) is feasible only if a small number of residues  $N$  are allowed to vary or if the number of amino acids  $m$  is greatly reduced. If, in the optimization process, the different sidechain conformations (rotamer states) of each amino acid are also considered (see [3]), the complexity of the search increases still further, because  $m$ , the number of possible 'states' per residue, increases by a factor of ten or more. Although complete enumeration is typically not feasible, sequence space can be sampled in a directed manner in order to find optimal (or nearly optimal) sequences. Stochastic methods, such as genetic algorithms or simulated annealing, involve searching sequence space in a partially random fashion; on average, the search progressively moves toward better scoring (lower energy) sequences [4,5]. The partially random nature of the search permits escape from local minima in the sequence/rotamer landscape. Using a simplified model, the Takada and Tamura groups have included information about unfolded structures (negative design) in a stochastic search for a sequence with a 'funneled conformational energy landscape' [6]. One

47-residue three-helix bundle protein so selected has CD and NMR spectral features of folded proteins (W Jin, O Kambara, H Sasakawa, A Tamura, S Takada, personal communication). When applied to atomically detailed representations, the stochastic methods focus primarily on repacking the interior of a structure with hydrophobic residues [7] and have been applied to the wild-type structures of 434 Cro [8], ubiquitin [9], the B1 domain of protein G [10], the WW domain [11] and helical bundles [11,12]. Although, in many cases, these methods have identified experimentally viable sequences [11,13], stochastic search methods need not identify global optima [14]. For potentials comprising only site and pair interactions, elimination methods such as 'dead end elimination' can find the global optimum [14,15-17]. Such methods successively remove individual amino acid rotamer states that cannot be part of the global optimum until no further states can be eliminated. The Mayo group applied such methods to automate the full sequence design of both a 28-residue zinc finger mimic [18] and, after predetermining hydrophobic and polar sites, a 51-residue homeodomain motif [19]. The group has also redesigned portions of a variety of proteins [20-22]. Functional properties such as metal binding or catalysis may also be included as elements of the design process [23,24]. The elements and algorithms of directed protein design have been the subject of several recent reviews [11,25,26].

Despite some striking successes, computational methods for directed design have limitations with respect to both identifying folding sequences and characterizing the features of protein sequences that share a common structure. Stochastic methods, such as simulated annealing or genetic algorithms, can be applied to large proteins and permit many sites to be varied simultaneously, but the computational times and resources required for such calculations are extensive, even for small proteins. When used as optimization methods, directed approaches will necessarily be sensitive to the energy or scoring function used. All energy functions in use in protein design, however, are necessarily approximate and uncertainties in the energy function may not merit the search for global optima. Furthermore, many naturally occurring proteins are not optimized. In fact, most proteins are only marginally stable (e.g.  $\Delta G^\circ < 10$  kcal/mol for folding) [27]. In addition, sequences that function, for example, those that bind another molecule, need not be the global optimum with respect to structural stability. Although stochastic methods can sample such suboptimal sequences, in general an exponentially large number of them will be possible and such sampling will be time consuming. Thus, it is important to develop methods complementary to those used for directed protein design — methods that reveal the features of sequences that are likely to fold into a particular structure but that may not be structurally 'optimal'. Such computational methods will have application to a new class of protein design studies, combinatorial experiments, in which large numbers of proteins may be simultaneously synthesized and screened.

## Combinatorial design

Combinatorial design provides a complementary approach to directed design for understanding sequence/structure compatibility and discovering novel sequences that fold into a specific structure. Combinatorial methods are powerful tools for cases in which we have an incomplete understanding of molecular properties. In protein combinatorial design experiments, large numbers of sequences (libraries) are screened for evidence of folding into a predetermined structure. A combinatorial experiment has two key elements: creating a library with a desired degree of diversity and assaying for sequences with 'protein-like' properties in terms of their structure or function. Depending upon how the diversity is generated and assayed, experiments of this type can explore a large number of sequences, up to  $10^{12}$  [28]. Certainly, such methods can be used to discover 'hits', that is, a few sequences that are especially stable or that are unusually strong in their function or binding properties. In addition, combinatorial experiments readily generate a sequence ensemble. Thus, using combinatorial experiments, we can potentially 'expand the protein sequence database' and the diversity of these additional sequences will be at the control of the researcher. Features important to folding (and other properties) may be explored in a way that is decoupled from the evolutionary requirements of nature's proteins. For example, these methods have been used to identify helical proteins [29-31], ubiquitin variants [32], self-assembled protein monolayers [33], proteins with amyloid-like properties [33], metal-binding peptides [34] and stable interhelical oligomers [35]. Several excellent reviews of combinatorial experiments have appeared recently [36,37,38,39].

The complexity of combinatorial experiments implies that limitations must be placed on the sequences, because the number that can be created and screened ( $10^6$ – $10^{12}$ ) is infinitesimal compared to the number possible (e.g.  $10^{130}$ ). Limitations on sequence properties are often guided by qualitative chemical considerations, but quantitative computational methods will be helpful in designing and interpreting combinatorial experiments.

The Hecht group has probed the extent to which the patterning of hydrophobic and hydrophilic residues can successfully reduce complexity in combinatorial design. While maintaining the periodicity of  $\alpha$  helices and  $\beta$  sheets in particular tertiary structures, such patterning is applied in order to expose hydrophilic residues to solvent and to sequester hydrophobic residues in the interior of the protein. Early targets were helical proteins; a fiducial 74-residue four-helix bundle was the template structure [40]. Such a structure has more than  $20^{74} \approx 10^6$  possible sequences. After binary patterning, five hydrophobic and six hydrophilic amino acids were permitted at 24 interior and 36 exterior positions, respectively, thus reducing the total number of possible sequences to  $10^{41}$ . From a protein library consistent with this binary patterning, a set of 50 correctly expressed sequences was selected for further

study. Around half of the 50 sequences isolated are protein-like in many respects [30], including their thermal denaturation [41]. About half the isolated sequences also bind heme [29] and many of these display carbon monoxide binding [42\*] or peroxidase activity [43]. This is surprising given that such functions were not part of the design or selection of the sequences. In a second-generation design, the group added six residues to each of the four helices of one of the most protein-like sequences. The additional residues were combinatorially patterned, as in the original experiment [39\*\*]. For these 102-residue sequences, the free energies of folding are increased 2–3-fold and the NMR data suggest well-determined structures. Using binary patterning of hydrophobicity consistent with an amphiphilic  $\beta$  sheet [44], the Hecht group has also identified proteins that aggregate to form amyloid fibrils [45] and crafted monomeric  $\beta$  proteins by introducing a nonpolar lysine mutation at the 'edge' strand of the target  $\beta$  sheet [46\*\*].

Despite the striking results from hydrophobic patterning, more detailed methods for library design are merited. Many of the hydrophobically patterned sequences that appear well structured are not sufficiently soluble for NMR structure determination [46\*\*] and, as a result, little is known concerning their structures at the atomic scale. Not all of the  $\alpha$ -helical sequences exhibit the sharp thermal transition seen in natural proteins (usually associated with a large  $\Delta H$  of folding). Such sequences may not possess well-packed interiors [41]. In natural proteins, the side-chains of most interior residues are well determined, as opposed to the variability that is obtained using hydrophobic patterning alone and that is observed in many *de novo* designed proteins [13,18]. A more fine-grained dictation of the amino acid identities is probably necessary for obtaining libraries that are rich in sequences with well-defined structures. Moreover, a more detailed specification of amino acid identities yields fewer sequences than hydrophobic patterning alone and further reduces the complexity of the library.

### Theories of combinatorial libraries

Surveying the complete sequence landscape of proteins seems, at first glance, intractable to both experiment and computation. In addition to the enormous number of possible sequences, many examples exist in nature of dissimilar sequences folding to essentially the same structure. Hence, sequence properties are nontrivial and proteins sharing a common structure can be nonlocal in sequence space. Nonetheless, computational methods permit us to estimate the properties, particularly the amino acid probabilities, of sequences consistent with a target structure.

Repeated use of directed search methods can estimate the properties of an ensemble of sequences. Desjarlais and co-workers have used independent runs of their sequence prediction algorithm across an ensemble of closely related structures all consistent with a particular fold (JR Desjarlais *et al.*, personal communication). For each

structure, an optimal 'nucleating' sequence is identified and subsequently the sequence/rotamer variability is explored throughout the structure. The method identifies effective reduced partition sums for each sequence/rotamer state and amino acid probabilities may be obtained at each residue position. The number of sequences decreases with stability, so the degree of complexity can be tuned by varying a cutoff in the effective free energies of the sequences. The method has been used to identify sequences consistent with the fold of a WW domain, a small  $\beta$ -sheet protein [1\*\*], some of which are currently being experimentally characterized.

The amino acid frequencies can also be determined directly, using a statistical theory of combinatorial libraries [47,48\*\*,49\*\*]. Ideas from statistical mechanics are used to address the number and composition of sequences that are consistent with a particular backbone structure. The theory addresses the whole space of available compositions, not just the small fraction that is accessible to experiment and to computational enumeration and sampling. The theory takes as input a target backbone structure and a scoring or energy function for quantifying sequence/structure compatibility. Global and local features can be prespecified using constraints on the sequences. For example, such constraints can be used to determine the energy the sequences assume in the target structure, the patterning of amino acids and the number of each amino acid present (composition). The theory yields estimates of both the number of sequences consistent with these constraints and the amino acid probabilities at each residue position. These residue-specific probabilities are the most probable such set and are determined — as in statistical mechanics — by maximizing an effective entropy, whereby this maximization is subject to constraints. Just as in thermodynamics, the judicious use of constraints can be used to reduce the entropy or the number of possible sequences. Thus, these methods provide a systematic means to focus the library, winnowing numbers such as  $10^{130}$  to numbers that are experimentally manageable, for example,  $10^6$ . The theory agrees well with exact results obtained with lattice models of proteins [47,48\*\*]. This method has been extended to realistic representations of proteins, in which the effects of sidechain packing are included in an atom-based manner [49\*\*]. The calculated sequence probabilities of the immunoglobulin light chain binding domain of protein L are in agreement with the frequencies observed in combinatorial phage display experiments [50,51]. These statistical methods have several advantages. They may be applied to much larger proteins ( $N > 100$  residues) and permit much larger sequence variation than many directed methods. They are sufficiently rapid that many backbone structures may be considered and those features that are robust with respect to minor structure modifications may be identified. Importantly, such methods provide perhaps the most natural input for a combinatorial experiment, the probabilities of the amino acids at each position among the sequences of a library. These amino acid

probabilities can also be used to identify specific amino acid sequences, which can then be synthesized; a consensus sequence comprising the most probable amino acid at each site can be selected or the probabilities can be used to bias a stochastic search for viable sequences (J Zou, JG Saven, unpublished data).

If the energy of the target state is one of the constraints, the statistical method reduces to an effective mean field theory. Mean field theories have seen extensive application in physical science and in biomolecular theory [52], and to protein evolution and natural sequence variability ([53]; H Kono, JG Saven, unpublished data). Voigt *et al.* [14\*] have compared mean field theories with directed search methods for identifying ground state sequence/rotamer combinations in protein design. They found that, although often more rapid, mean field theories do not always identify such ground states. Interestingly, Voigt *et al.* applied the mean field theory to large proteins (subtilisin E and T4 lysozyme) to determine local site entropies,  $s_p$ , where  $\exp(s_p)$  quantifies the effective number of amino acids allowed at residue  $i$  in a structure [54\*\*,55]. Sites with large values of  $s_p$ , those most tolerant to mutation [56], are likely to support substitutions that improve stability or function when *in vitro* evolution experiments are used to explore sequence space [37]. For such experiments, the mutation rate is low enough that multiple mutations of strongly interacting sites are rare. Thus, mutations that improve 'fitness' are most likely to accumulate at sites that are the most 'decoupled' from other sites. Such mutations can potentially be targeted for variation in an *in vitro* evolution experiment.

## Conclusions

Much recent progress has been seen in the design and discovery of new proteins, and combinatorial approaches are accelerating the pace. Such methods are most useful when our quantitative understanding of important protein properties, such as stability and catalytic activity, is limited. Not only can combinatorial methods be used for discovery but also, more deeply, they can inform our understanding of protein properties by generating and assaying whole ensembles of sequences. Traditionally, advances in structural biology have come from examining the structures of naturally occurring proteins, but, with combinatorial experiments, an enormous diversity of sequences can be generated at the control of the researcher. Detailed questions can be addressed, such as the utility of hydrophobic patterning or of predetermining particular sites for amino acid variation. Theory and simulation will continue to aid the design and interpretation of combinatorial experiments. Such methods will also facilitate the exploration of what is possible with the amino acids: how diverse is the set of all possible sequences that fold to a particular structure and what structures not yet seen in nature can be crafted with the amino acids? Such methods will perhaps have an even more profound impact on designing nonbiological foldamers [57\*\*], structures about which we have much less empirical information than we do about biopolymers.

## Acknowledgements

The author acknowledges support from the National Science Foundation (CHE 98-16497 and CHE 99-84752). JGS is a Cottrell Scholar of Research Corporation and is an Arnold and Mabel Beckman Foundation Young Investigator.

## References and recommended reading

Papers of particular interest, published within the annual period of review, have been highlighted as:

- of special interest
  - of outstanding interest
1. Kraemer-Pecore CM, Wolcott AM, Desjarlais JR: Computational •• protein design. *Curr Opin Chem Biol* 2001, 5:690-695. This is a compact but excellent review on recent progress in computational methods for protein design. The authors also discuss recent efforts in designing the WW domain, a small  $\beta$  protein.
  2. Bryson JW, Betz SF, Lu HS, Suich DJ, Zhou HX, O'Neill KT, DeGrado WF: Protein design: a hierarchical approach. *Science* 1995, 270:935-941.
  3. Dunbrack R: Rotamer libraries. *Curr Opin Struct Biol* 2002, 12:in press.
  4. Shakhnovich EI, Gutin AM: A new approach to the design of stable proteins. *Protein Eng* 1993, 6:793-800.
  5. Jones DT: *De novo* protein design using pairwise potentials and a genetic algorithm. *Protein Sci* 1994, 3:567-574.
  6. Onuchic JN, Luthey-Schulten Z, Wolynes PG: Theory of protein folding: the energy landscape perspective. *Annu Rev Phys Chem* 1997, 48:545-600.
  7. Hellings HW, Richards FM: Optimal sequence selection in proteins of known structure by simulated evolution. *Proc Natl Acad Sci USA* 1994, 91:5803-5807.
  8. Desjarlais JR, Handel TM: *De-novo* design of the hydrophobic cores of proteins. *Protein Sci* 1995, 4:2006-2018.
  9. Johnson EC, Lazar GA, Desjarlais JR, Handel TM: Solution structure and dynamics of a designed hydrophobic core variant of ubiquitin. *Structure* 1999, 7:967-976.
  10. Jiang X, Farid H, Pistor E, Farid RS: A new approach to the design • of uniquely folded thermally stable proteins. *Protein Sci* 2000, 9:403-416. The authors use a novel scoring function for the design of hydrophobic interiors. In addition to steric interactions, the function includes parameterizations of changes in the heat capacity and the conformational entropy upon folding. Simulated annealing was used to optimize the score. The backbone and exterior residue identities were constrained. In tests on two small proteins, in which 10-11 interior residues were varied, the native sequence was regenerated, as well as the sequences of known stable variants. Interestingly, previously designed sequences with low stability and weak cooperativity were not identified. In larger proteins tested, in which 32 and 63 residues were varied, sequence/rotamer combinations close to native were identified.
  11. Jiang X, Bishop EJ, Farid RS: A *de novo* designed protein with properties that characterize natural hyperthermophilic proteins. *J Am Chem Soc* 1997, 119:838-839.
  12. Bryson JW, Desjarlais JR, Handel TM, DeGrado WF: From coiled coils to small globular proteins: design of a native-like three-helix bundle. *Protein Sci* 1998, 7:1404-1414.
  13. Walsh STR, Cheng H, Bryson JW, Roder H, DeGrado WF: Solution structure and dynamics of a *de novo* designed three-helix bundle protein. *Proc Natl Acad Sci USA* 1999, 96:5486-5491.
  14. Voigt CA, Gordon DB, Mayo SL: Trading accuracy for speed: • a quantitative comparison of search algorithms in protein sequence design. *J Mol Biol* 2000, 299:789-803. The authors carefully compared different methods for sequence design, including simulated annealing, genetic algorithms, mean field methods and dead end elimination (DEE). DEE most reliably finds global minima, but the authors also note that the method may be limited to 30 amino acid sites for which full amino acid variability is permitted. The authors extrapolate the results to regimes to which DEE cannot be applied. They find that both mean field and annealing approaches perform best with core residues and less reliably with residues that are fully or partially solvent exposed.
  15. Gordon DB, Mayo SL: Radical performance enhancements for combinatorial optimization algorithms based on the dead-end elimination theorem. *J Comput Chem* 1998, 19:1505-1514.

16. Gordon DB, Mayo SL: Branch-and-terminate: a combinatorial optimization algorithm for protein design. *Structure* 1999, 7:1089-1098.
17. Pierce NA, Spriet JA, Desmet J, Mayo SL: Conformational splitting: a more powerful criterion for dead-end elimination. *J Comput Chem* 2000, 21:999-1009.
18. Dahiyat BI, Mayo SL: *De novo* protein design: fully automated sequence selection. *Science* 1997, 278:82-87.
19. Marshall SA, Mayo SL: Achieving stability and conformational specificity in designed proteins via binary patterning. *J Mol Biol* 2001, 305:619-631.
- The authors combined binary patterning with atomistic sidechain interactions to identify folding sequences of a homeodomain. Preliminary calculations were done using 'generic amino acids' to identify those sites that are most appropriate for polar or hydrophobic residues. Subject to this binary patterning, a directed search was then performed using dead end elimination. Interestingly, the authors identified an optimal binary patterning, whereby adding or subtracting hydrophobic residues adversely affects folding to stable monomers.
20. Malakauskas SM, Mayo SL: Design, structure, and stability of a hyperthermophilic protein variant. *Nat Struct Biol* 1998, 5:470-475.
21. Strop P, Mayo SL: Rubredoxin variant folds without iron. *J Am Chem Soc* 1999, 121:2341-2345.
22. Shimaoka M, Shifman JM, Jing H, Takagi L, Mayo SL, Springer TA: Computational design of an Integrin I domain stabilized in the open high affinity conformation. *Nat Struct Biol* 2000, 7:674-678.
23. DeGrado WF, Summa CM, Pavone V, Nastri F, Lombardi A: *De novo* design and structural characterization of proteins and metalloproteins. *Annu Rev Biochem* 1999, 68:779-819.
24. Bolon DN, Mayo SL: Enzyme-like proteins by computational design. *Proc Natl Acad Sci USA* 2001, 98:14274-14279.
- The authors designed non-native enzymatic activity into a thioredoxin fold. The authors computationally identified promising active sites on the scaffold. The sequence was designed to stabilize the transition state of a hydrolysis reaction. The enzymes so designed had activity well above background.
25. Street AG, Mayo SL: Computational protein design. *Structure* 1999, 7:R105-R109.
26. Saven JG: Designing protein energy landscapes. *Chem Rev* 2001, 101:3113-3130.
- The author reviews recent progress in protein design from the perspective of the energy landscape theory of folding. In the context of theory, models and real systems, different issues involved in design are discussed, including target structures, energy functions, foldability criteria, search methods and the size of the amino acid alphabet.
27. Gromiha MM, Uedaira H, An J, Selvaraj S, Prabakaran P, Sarai A: ProTherm, thermodynamic database for proteins and mutants: developments in version 3.0. *Nucleic Acids Res* 2002, 30:301-302.
28. Keefe AD, Szostak JW: Functional proteins from a random-sequence library. *Nature* 2001, 410:715-718.
- A fascinating study on a random search for 'function' among random amino acid sequences. Using combinatorial methods the authors have pioneered, ATP-binding proteins were selected from a library of 10<sup>12</sup> sequences.
29. Rojas NRL, Kamtekar S, Simons CT, Mclean JE, Vogel KM, Spiro TG, Farid RS, Hecht MH: *De novo* heme proteins from designed combinatorial libraries. *Protein Sci* 1997, 6:2512-2524.
30. Roy S, Ratnaswamy G, Bolce JA, Fairman R, McLendon G, Hecht MH: A protein designed by binary patterning of polar and nonpolar amino acids displays native-like properties. *J Am Chem Soc* 1997, 119:5302-5306.
31. Roy S, Helmer KJ, Hecht MH: Detecting native-like properties in combinatorial libraries of *de novo* proteins. *Fold Des* 1997, 2:89-92.
32. Finucane MD, Tuna M, Lees JH, Woodson DN: Core-directed protein design. I. An experimental method for selecting stable proteins from combinatorial libraries. *Biochemistry* 1999, 38:11604-11612.
33. Xu GF, Wang WX, Groves JT, Hecht MH: Self-assembled monolayers from a designed combinatorial library of *de novo* beta-sheet proteins. *Proc Natl Acad Sci USA* 2001, 98:3652-3657.
34. Case MA, McLendon GL: A virtual library approach to investigate protein folding and internal packing. *J Am Chem Soc* 2000, 122:8089-8090.
35. Arndt KM, Pelletier JN, Muller KM, Alber T, Michnick SW, Pluckthun A: A heterodimeric coiled-coil peptide pair selected *in vivo* from a designed library-versus-library ensemble. *J Mol Biol* 2000, 295:627-639.
36. Zhao HM, Arnold FH: Combinatorial protein design: strategies for screening protein libraries. *Curr Opin Struct Biol* 1997, 7:480-485.
37. Giver L, Arnold FH: Combinatorial protein design by *in vitro* recombination. *Curr Opin Chem Biol* 1998, 2:335-338.
38. Hoess RH: Protein design and phage display. *Chem Rev* 2001, 101:3205-3218.
- A comprehensive review of a commonly used method to generate and display combinatorial libraries of proteins and peptides.
39. Moffet DA, Hecht MH: *De novo* proteins from combinatorial libraries. *Chem Rev* 2001, 101:3191-3203.
- A review of recent work on the *de novo* combinatorial design of proteins, focusing primarily on the pioneering work of the Hecht group.
40. Kamtekar S, Schiffer JM, Xiong HY, Babik JM, Hecht MH: Protein design by binary patterning of polar and nonpolar amino-acids. *Science* 1993, 262:1680-1685.
41. Roy S, Hecht MH: Cooperative thermal denaturation of proteins designed by binary patterning of polar and nonpolar amino acids. *Biochemistry* 2000, 39:4603-4607.
42. Moffet DA, Case MA, House JC, Vogel K, Williams RD, Spiro TG, McLendon GL, Hecht MH: Carbon monoxide binding by *de novo* heme proteins derived from designed combinatorial libraries. *J Am Chem Soc* 2001, 123:2109-2115.
- Heme-assisted binding of a diatomic ligand turns out to be easier to find than expected within a library of sequences patterned to form a four-helix bundle. Eight combinatorially selected heme-binding sequences bind carbon monoxide with an affinity similar to that of myoglobin. The binding properties of the proteins aren't nearly as diverse as those seen among natural heme proteins, but these *de novo* sequences serve as a useful 'reference'.
43. Moffet DA, Certain LK, Smith AJ, Kessel AJ, Beckwith KA, Hecht MH: Peroxidase activity in heme proteins derived from a designed combinatorial library. *J Am Chem Soc* 2000, 122:7612-7613.
44. West MW, Beasley JR, Hecht MH: Collections of *de novo* beta-sheet proteins designed by binary patterning of polar and nonpolar amino acids. *Protein Eng* 1997, 10:38-38.
45. West MW, Wang WX, Patterson J, Mancias JD, Beasley JR, Hecht MH: *De novo* amyloid proteins from designed combinatorial libraries. *Proc Natl Acad Sci USA* 1999, 96:11211-11216.
46. Wang WX, Hecht MH: Rationally designed mutations convert *de novo* amyloid-like fibrils into monomeric beta-sheet proteins. *Proc Natl Acad Sci USA* 2002, 99:2760-2765.
- Previously, the authors had used hydrophobic patterning consistent with  $\beta$  sheets that intermolecularly align 'edge on' and found these sequences did indeed form the amyloid fibrils that were expected. In this paper, they break up these edge-on interactions with a hydrophilic residue (lysine) at each edge of the  $\beta$  sheet. These sequences are indeed monomeric and appear to be well structured according to CD and NMR peak dispersion. These are the first examples of combinatorial  $\beta$ -protein design.
47. Saven JG, Wolynes PG: Statistical mechanics of the combinatorial synthesis and analysis of folding macromolecules. *J Phys Chem B* 1997, 101:8375-8389.
48. Zou JM, Saven JG: Statistical theory of combinatorial libraries of folding proteins: energetic discrimination of a target structure. *J Mol Biol* 2000, 296:281-294.
- The authors extend their statistical theory of sequence libraries to include negative design. The sequence space is resolved in multiple dimensions and the number of sequences is characterized according to the folded state energy and stability gap (the difference in energy between the folded state and an ensemble of unfolded conformations). Excellent agreement is observed between theoretical and exact lattice model results for both the numbers of sequences and the monomer probabilities.
49. Kono H, Saven JG: Statistical theory for protein combinatorial libraries. Packing interactions, backbone flexibility, and the sequence variability of a main-chain structure. *J Mol Biol* 2001, 306:607-628.
- A statistical theory of combinatorial libraries developed for combinatorial experiments. The authors used an atom-based potential and rotamer states to identify the sequence probabilities consistent with a particular structure. An effective one-body energy was introduced that relates the hydrophobicity

or solvent-exposure propensity to local  $\beta$ -carbon density. The calculations give good results with regard to sidechain modeling. Calculations were done that are consistent with recent combinatorial experiments on protein L. Generally, the calculations are in good agreement with the observed amino acid frequencies, despite the sampling issues that are always a concern with these comparisons. (Only 20–40 sequences were sequenced in the experiments.)

50. Kim DE, Gu HD, Baker D: The sequences of small proteins are not extensively optimized for rapid folding by natural selection. *Proc Natl Acad Sci USA* 1998, 95:4982-4986.
51. Gu H, Doshi N, Kim DE, Simons KT, Santiago JV, Nauli S, Baker D: Robustness of protein folding kinetics to surface hydrophobic substitutions. *Protein Sci* 1999, 8:2734-2741.
52. Koehl P, Delarue M: Mean-field minimization methods for biological macromolecules. *Curr Opin Struct Biol* 1996, 6:222-226.
53. Dokhotyan NV, Shakhnovich EI: Understanding hierarchical protein evolution from first principles. *J Mol Biol* 2001, 312:289-307.
54. Voigt CA, Mayo SL, Arnold FH, Wang ZG: Computational method •• to reduce the search space for directed protein evolution. *Proc Natl Acad Sci USA* 2001, 98:3778-3783.  
The authors used a mean field theory to determine each residue's structural tolerance to mutations. This tolerance is quantified by the residue's local sequence entropy, which is a measure of the effective number of amino acids that are structurally permitted at that site. For an *in vitro* directed evolution experiment, the authors suggest that mutations that enhance stability or activity are most likely to accumulate in these high entropy regions. Multiple compensating mutations are rare in such experiments, so mutations are most likely at sites that tolerate multiple amino acids. Calculations involving subtilisin E and T4 lysozyme are consistent with the mutations observed in directed evolution experiments.
55. Voigt CA, Mayo SL, Arnold FH, Wang ZG: Computationally focusing the directed evolution of proteins. *J Cell Biochem* 2001:58-63.
56. Sander C, Schneider R: Database of homology-derived protein structures and the structural meaning of sequence alignment. *Proteins* 1991, 9:56-68.
57. Hill DJ, Mio MJ, Prince RB, Hughes TS, Moore JS: A field guide to •• foldamers. *Chem Rev* 2001, 101:3893-4011.  
A comprehensive review of nonbiological folding molecules.